Preprints of the 21st IFAC World Congress (Virtual)
Berlin, Germany, July 12-17, 2020

K-Means Clustering-based Kernel Canonical Correlation Analysis for Multimodal Emotion Recognition ⋆

Luefeng Chen [*,**], **Kuanlin Wang** [*,**], **Min Wu** [*,**,†], **Witold Pedrycz** [***], **Kaoru Hirota** [****]

[*] *School of Automation, China University of Geosciences, Wuhan 430074, China (e-mail: chenluefeng@cug.edu.cn)*
[**] *Hubei Key Laboratory of Advanced Control and Intelligent Automation for Complex Systems, Wuhan 430074, China (e-mail: wumin@cug.edu.cn)*
[***] *Department of Electrical and Computer Engineering, University of Alberta, Edmonton, AB T6R 2G7, Canada (e-mail: wpedrycz@ualberta.ca)*
[****] *Tokyo Institute of Technology, Yokohama 226-8502, Japan (e-mail: hirota@jsps.org.cn)*

Abstract: Emotion is an important part of human interaction. Emotional recognition can greatly promote human-centered interaction techniques. On this basis, multimodal feature fusion can effectively improve the emotion recognition rate. However, in the multimodal feature fusion at the feature level, most of the methods do not consider the intrinsic relationship between different modes. Only the fusion of analysis and transformation of the feature matrices of different modes does not make better use of modal differences to improve the recognition rate. This problem led us to propose feature fusion method based on K-Means clustering and kernel canonical correlation analysis (KCCA). Clustering makes the classification of features not classified by mode, but by the degree of influence on emotional labels, thus positively affecting the results of KCCA. The experimental results obtained on the Savee database show that the proposed K-Means based KCCA improves overall classification performance and produces higher recognition rate than that of the state of art methods, such as the Informed Segmentation and Labeling Approach.

Keywords: Emotion Recognition, K-Means Clustering, Kernel Canonical Correlation Analysis, Feature Fusion

1. INTRODUCTION

Human-computer interaction is a research hotspot in the field of Artificial Intelligence [Chen et al. (2018)]. It uses signals such as video and audio to affect the physiological, posture, expression and speech caused by human emotions [Chen et al. (2019)]. However, due to the nature of emotion, there are many limitations, such as missing and poor quality feature extraction, so that we cannot observe the ideal characteristics. In contrast, multimodal fusion can complement existing technique, thereby improving the accuracy of information estimation. Feature-level fusion methods have been applied to the feature extraction and led to better results. It is to combine two sets of different features of the same sample and combine them into a newly generated feature space by PCA [Moore (1981)].

⋆ This work was supported by the National Natural Science Foundation of China under Grants 61973286, 61603356, and 61733016, the 111 project under Grant B17040, and the Fundamental Research Funds for the Central Universities, China University of Geosciences (No. 201839).
† Corresponding author: Min Wu (e-mail: wumin@cug.edu.cn).

In the existing feature fusion methods, a method called serial fusion [Liu and Wechsler (2001)]. Canonical Correlation Analysis (CCA) [Hotelling (1935)] is a fusion method whose purpose is to find a pair of projection directions so that there is a maximum correlation between the two sets of features. When using this method to cope with nonlinear problems, the problem of under-learning will inevitably occur. To make up for this deficiency, the method of kernels has been applied [Melzer et al. (2003)]. Then proposed was Kernel Canonical Correlation Analysis, abbreviated as KCCA, which uses kernel techniques to linearly extract nonlinear features. In general, we call it a double canonical correlation analysis.

In recent years, many scholars have proposed the theory of Multi-set Canonical Correlation Analysis (MCCA) [Nielsen (2002)]. It is suitable for multi-feature fusion in a mode and has been used to combine finger vein, fingerprint, finger shape and finger knuckle print features of a single human finger for finger biometrics [Peng et al. (2015)]. A comparison of decision-level and feature-level fusions was mentioned [Planet and Sanz (2012)]. Decision-

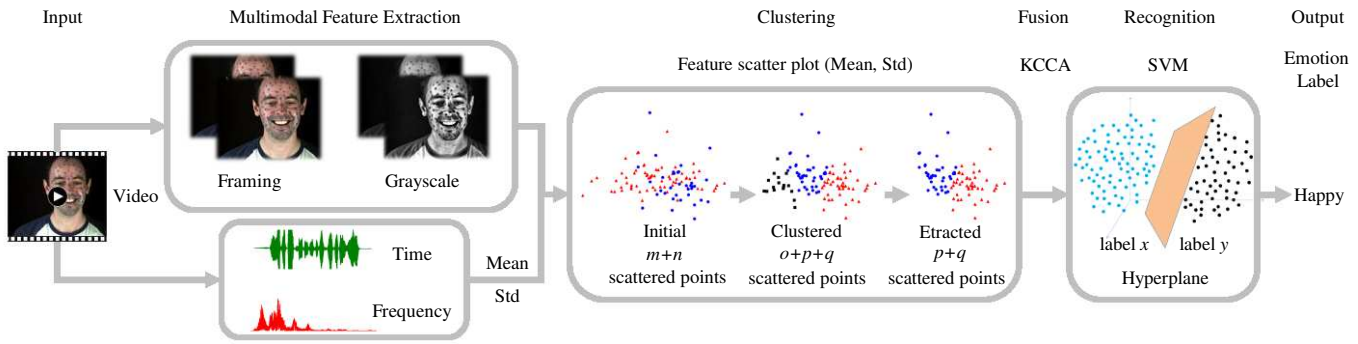

Fig. 1. An overview of our presented audio-visual emotion recognition system based on feature fusion. The difference between the modes makes the dimensions of the feature matrix different. Here, the $m+n$ dimensional samples are clustered into 3 classes, which are respectively o, p, q. And the feature class o is extracted according to the standard.

based classifier fusion is also applicable to a single mode [Chen et al. (2020)]. For the K-Means clustering algorithm, it is often used in emotion recognition, and it has achievements in the feature extraction of speech [Sultana and Shahnaz (2014)] and expression [Ghahari et al. (2009)].

2. K-MEANS CLUSTERING BASED KCCA

This section adopts emotional feature extraction based on speech and facial expressions and fusion algorithm based on the feature level. The overall algorithm is shown in Fig. 1. For sentiment samples of speech, we extract its comprehensive speech emotion features (including time domain, frequency domain, MFCCs, etc.); for the emotion samples of facial expression, one extracts its pixel features based on grayscale equalization. The multimodal features are reclassified using K-Means clustering and then based on the feature level fusion method of kernel canonical correlation analysis to extract features, determine parameters and get the best recognition rate by SVM.

2.1 Preprocessing

For facial expression data, firstly, we extract 30 frames of images by equal frame distance according to the total number of frames per video. The key area of each facial image is extracted based on Viola-Jones algorithm, then images are normalized to a uniform scale. The preprocessed facial expression image frame is obtained; for each speech signal, the endpoint detection is performed, the blank frame segment is deleted, and they are divided into equal distance frames.

2.2 Feature Extraction of Speech

Compared with the image features of facial expressions, the extraction of emotional features of speech information is relatively complicated, but main ideas and methods in speech feature extraction have gradually matured, and there is a good accuracy obtained in the field of emotion recognition. In this paper, the temporal domain features (feature 1-3), frequency domain features (feature 4-8), Mel cepstrum coefficients (features 9-21) and related temperament features (features 22-34) are used in speech emotion feature extraction. These 34-dimensional feature vectors were used in the extraction process.

In the field of sound processing, Mel-Frequency Cepstrum is a linear transformation based on the nonlinear logarithmic energy spectrum. Mel-Frequency Cepstral Coefficients (MFCCs) constitute the frequency spectrum of Mel frequency. The difference between cepstrum and Mel frequency cepstrum is that the frequency division of the Mel frequency cepstrum is equally spaced on the Mel scale. It is more similar to the human auditory system than the linearly spaced bands used in normal cepstrum. The nonlinear representation can better represent acoustic signals in several areas.

2.3 Feature Extraction of Expression

The advantage of the pixel method is that the feature information is complete. Its disadvantage is that the feature dimensionality is high and is affected by the pose. Combining the selected facial expression database, the pixel information is processed by the pixel method, and the relevant features are extracted for the purpose of identification. Since the color of a pixel is represented by three RGB values, the pixel matrix corresponds to an array of three color vectors. For color images, the grayscale value is obtained by function mapping. The gray scale range is usually in the range of 0 to 255. The higher the gray value, the brighter the pixel is. The gray level of the image is such that each pixel in the matrix of pixels satisfies the equal relationship, which is called the gray value.

$$Gray = R * 0.3 + G * 0.59 + B * 0.11 \tag{1}$$

Adaptive Histogram Equalization (AHE) improves image contrast. The AHE algorithm calculates the local histogram of the image and then redistributes the brightness to change the contrast of the image. The algorithm improves the local contrast of the image and is suitable for obtaining more details of the image.

2.4 K-Means Clustering of features

Before proceeding with the KCCA fusion, features of facial expressions and speech are based on a clustering method, namely K-means. K-means is a widely used unsupervised learning algorithm. The goal we need to achieve is just to reveal a strucure in data. For a given sample set, it is divided into K clusters based on the distance between the samples. The K-means clustering method consists of the following steps:

Step 1: Divide the data set $x_1, x_2, ..., x_m$ into k clusters.

Step 2: Randomly select k cluster centroids as $\mu_1, \mu_2, ..., \mu_k$.

Step 3: Repeat the process until its convergence:

(1) For each x_i, calculate the cluster it should belong to

$$c_i = \arg\min_j \|x_i - \mu_j\|^2 \tag{2}$$

(2) For each cluster j, recalculate the centroid of the class

$$\mu_j = \frac{\sum_{i=1}^{m} 1\{c_i = j\} x_i}{\sum_{i=1}^{m} 1\{c_i = j\}} \tag{3}$$

where, K is the number of clusters given in advance. c_i represents the class closest to the x_i in the k classes, and the value of c_i is one of 1 to k. Center of gravity μ_j is the sample center point belonging to the same category.

2.5 Canonical Correlation Analysis

Assuming there are two sets of one-dimensional data X and Y, the correlation coefficient ρ is defined as

$$\rho(X,Y) = \frac{\mathrm{cov}(X,Y)}{\sqrt{D(X)}\sqrt{D(Y)}} \tag{4}$$

where $\mathrm{cov}(X,\ Y)$ is the covariance of X and Y, and $D(X)$, $D(Y)$ are the variances of X and Y, respectively. Although the correlation coefficient can help us analyse the correlation between one-dimensional data, it cannot be used directly for high-dimensional data. CCA provides a workaround.

$$X' = a^T X, Y' = b^T Y \tag{5}$$

The goal of CCA is to maximize $\rho(X', Y')$ and determine the corresponding projection vectors a, b, namely

$$\underbrace{\arg\max}_{a,b} \frac{\mathrm{cov}\left(X', Y'\right)}{\sqrt{D\left(X'\right)}\sqrt{D\left(Y'\right)}} \tag{6}$$

We can use the optimization method similar to SVM, fix the denominator, optimize the numerator, and the specific conversion is

$$\begin{aligned} &\underbrace{\arg\max}_{a,b} a^T S_{XY} b \\ &\text{s.t.} a^T S_{XX} a = 1, b^T S_{YY} b = 1 \end{aligned} \tag{7}$$

As long as the maximal value of the optimization target is obtained, it is the correlation measure of the multidimensional X and Y mentioned above, and the corresponding a and b are projection vectors when realizing dimensionality reduction. There are two methods of this optimization. The first one is the singular value decomposition SVD, and the second is the traditional Lagrangian feature decomposition. The results produceding these methods are the same.

2.6 Kernel Function

The main idea of the kernel method is feature mapping, which maps linear wording data in low-dimensional space to high-dimensional space called linear separable data in high-dimensional space. In this process, it is not necessary to know the specific form of the mapping function. It is necessary to realize the relevant conditions to meet the mapping requirements of the kernel function, and then optimize high-dimensional space so that the nonlinear data are linearly separable. Some kernel functions are shown as follows,

(1) Linear $K\left(x_1, x_2\right) =< x_1, x_2 >$ As one of the simplest and more common kernel functions, it is used where the sample dimensionality are not high and can be linearly separated in low dimensional spaces.

(2) Polynomial $K\left(x_1, x_2\right) = (< x_1, x_2 > +R)^p$ The polynomial kernel function can better describe the data, but it is also prone to over-fitting.

(3) Radial Basis Function $K\left(x_1, x_2\right) = e^{-\frac{\|x_1 - x_2\|^2}{2\sigma^2}}$ σ is the width of the RBF kernel. If $K\left(x_1, x_2\right)$ is small, the weight of the high-order feature attenuates slowly.

(4) Sigmoid $K\left(x_1, x_2\right) = \tanh\left(\alpha x^t y + c\right)$

(5) Spline $K\left(x_1, x_2\right) = 1 + x^t y + x^t y \min(x, y) - \frac{x+y}{2} \times \min(x, y)^2 + \frac{1}{3}\min(x, y)^3$

However, ordinary linear CCA can only explore the linear relationship. In reality, the relationship between variables is often nonlinear. KCCA brings the idea of kernel function into CCA. The following two sets of data, X and Y are projected to the high-dimensional space via the kernel function, and then the process is similar to CCA, and is projected again.

$$\begin{aligned} \psi(X) &= \alpha_{\phi(X)}^T \phi(X) \\ \psi(Y) &= \beta_{\phi(Y)}^T \phi(Y) \end{aligned} \tag{8}$$

$\alpha_{\phi(X)}$ and $\beta_{\phi(Y)}$ are solved, and the correlation between $\psi(X)$ and $\psi(Y)$ is maximized. According to the definition of the kernel function, the calculations of the kernel are as follows

$$\begin{aligned} K_X &=< \phi(X), \phi(X) >= \phi^T(X)\phi(X) \\ K_Y &=< \phi(Y), \phi(Y) >= \phi^T(Y)\phi(Y) \end{aligned} \tag{9}$$

Let $\alpha_{\phi(X)} = \phi(X)a$, $\beta_{\phi(Y)} = \phi(Y)b$. Then

$$\begin{aligned} \text{s.t. } &\alpha_{\phi(x)}^T \phi^T(X)\phi(X)\alpha_{\phi(X)} = 1 \\ &\beta_{\phi(Y)}^T \phi^T(Y)\phi(Y)\beta_{\phi(Y)} = 1 \end{aligned} \tag{10}$$

We obtain

$$\begin{aligned} &\alpha^T \phi^T(X)\phi(X)\phi^T(X)\phi(X)a = 1 \\ &\beta^T \phi^T(Y)\phi(Y)\phi^T(Y)\phi(Y)b = 1 \\ &a^T K_X K_X a = 1, b^T K_Y K_Y b = 1 \end{aligned} \tag{11}$$

In this case, it converts into an optimization problem. By using Lagrange multiplier, the optimal value is solved for the above problem, that is, the coefficients a and b are obtained, and the projection matrix can also be obtained.

2.7 Support Vector Machine

The purpose of the support vector machine is to find the best hyperplane through given positive and negative data. The concept of SVM is based on the two-category problem of linear conditions. H is the classification hyperplane of these two types of data, H_1, H_2 are the two hyperplanes closest to H, which are parallel to the classification plane.

The data point is represented by x, y is the category, ω is the hyperplane normal vector, and b is the deviation. This hyper-plane for i = 1, 2, ..., n. is described as

$$x = x_0 + r\frac{\omega}{\|\omega\|} \tag{12}$$

The distance from the data point to the hyperplane is called the geometric interval, denoted by

$$\begin{cases} \omega^T x_i + b > 0, y_i = 1 \\ \omega^T x_i + b < 0, y_i = -1 \\ \omega^T x_i + b = 0, y_i \neq \pm 1 \end{cases} \quad (13)$$

Where x_0 is positioned on the hyperplane, satisfying $\omega^T x + b = 0$, it is the geometric interval as

$$r = \frac{\omega^T x + b}{\|\omega\|} \quad (14)$$

When the data point and the hyperplane are separated by a large interval, the classification confidence is high, thereby defining the objective function of the maximum interval classifier as

$$\min \frac{1}{2}\|\omega\|^2, \text{s.t.} y_i(\omega^T x_i + b) \geq 1, i = 1, 2, \ldots, n \quad (15)$$

3. EXPERIMENTS ON K-MEANS BASED KCCA

SAVEE University's public multimodal database contains seven emotions, which are anger, disgusted, fear, happy, neutral, sad, and surprised. The SAVEE database records emotional video and audio from four British male speakers, all of which were recorded by graduate students and researchers at the University of Surrey. It contains a total of 480 sets of video and audio by the label. To extract facial expression features, the actor's front face was painted with 60 markers, which were painted on the forehead, eyebrows, cheeks, lips and chin. Each audio sample rate is 44.1 kHz and the duration is 2-5 s. Each emotion contains 15 text materials. Neutral emotion provides 15 other materials.

Fig. 2. Part of the sample frames in the SAVEE database.

For the feature matrix of the acquired facial grayscale, we obtain different results of emotion recognition by PCA feature extraction, as shown in Fig. 3, the red line represents the recognition rate, and the blue line represents the per-K main component. From the figure, we can see that when the facial feature is retained to 30-40 dimensions, the peak is reached. When we take 34-dimension, we reach the peak value, that is, the optimal recognition rate of facial expression emotion recognition equal to 88.89%.

We get a speech dimensionality of 68 and a facial expression dimensionality of 34. There is a nonlinear relationship between the two types of data, so we use the kernel method to raise the face expression to 68 variables and fuse the two types of data in series. According to different optimal recognition rates based on KCCA, the results of the kernel functions of Line, Sigmoid, Spline, RBF (gamma=0.01) and Polynomial (p=2) are 66.53%, 67.50%, 70.28%, 88.33% and 91.39%. It can be seen that the polynomial kernel is suitable for the identification of such data features, so we adjust the vaule of p. After

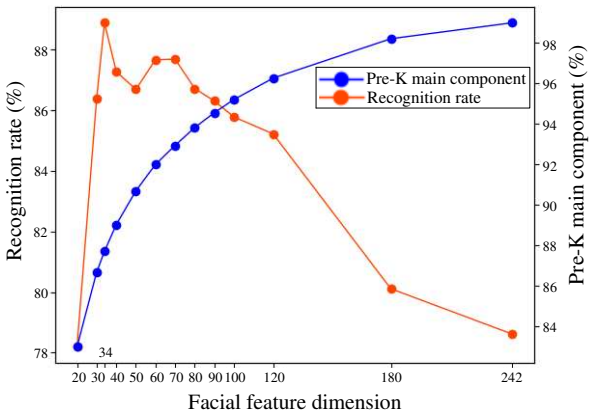

Fig. 3. Per-K main component and recognition rate under different face feature dimension.

the increase of p, the data after the dimension increase is too scattered, and the recognition rate decreases linearly. Therefore, based on the KCCA fusion method, the optimal recognition rate is 91.39%.

For the general treatment of the two types of data by KCCA fusion, and do not fully consider the intrinsic relationship between the features, so we consider the two types of features as the whole and reclassify. First, we normalize the dimensional data to map it to the (0, 1) interval, and then take the mean and standard deviation to form two-dimensional data. As shown in Fig. 4, we can see the two-dimensional data points in the graph. Blue dots represent facial expression features, and red dots represent speech features.

Next, we use K-Means for clustering, where data are divided into 3 categories, and iterates until each point is constant within a certain range from the centre of the cluster. As shown in Fig. 5, we get 3 types of data. We believe that the feature closer to the origin is considered to have a small influence on the emotion recognition result. Then, when we fuse, we discard the class represented by the black square.

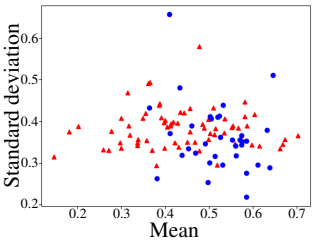

Fig. 4. Scatter plot composed of mean and standard deviation of two types of feature.

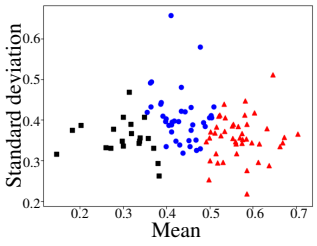

Fig. 5. Scatter plots of three types of features after K-Means clustering.

It is known that the red data are 44-dimensional, and the 38-dimensional blue data are raised to 44-dimensional and then CCA fusion is performed to obtain an 88-dimensional feature matrix, and 93.06% is recognized by the SVM Classifier.

The confusion matrix of facial emotion recognition results of K-Means + KCCA Fusion is shown in Fig. 6. It can

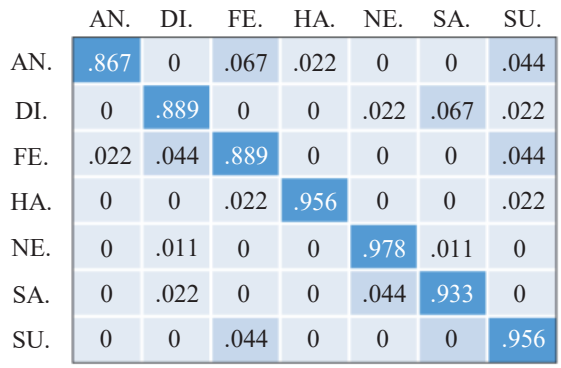

Fig. 6. Confusion matrix for SAVEE database.

be seen that the recognition rate of neutral emotions is higher. This may be due to the large difference between the speech of neutral emotions and the information about other emotions. Under the interaction of speech and facial expression, our presented method improves the recognition rate.

Table 1. Recognition rate (%) and feature dimensionality of features obtained under different methods.

Method	Recognition rate	Dimension
Face	88.89	34
Speech	59.72	68
Series Fusion	84.44	102
PCA + CCA Fusion	89.75	68
Kernel + CCA Fusion	91.39	136
K-Means + KCCA Fusion	93.06	88

We compare the recognition results with the audio-visual emotion recognition extracted by Kim and Provost (2019) based on Audio and Upper Face Region by using the Informed Segmentation and Labeling Approach (ISLA).

Table 2. Recognition rate (%) of existing method comparative analysis.

Recognition method	Face	Speech	Fusion
ISLA [Kim and Provost (2019)]	83.96	80.75	86.01
K-Means + KCCA	88.89	59.72	93.06

As shown in (2), we can see that our fusion method has a higher recognition rate based on the lower recognition rate of the single speech modality, which can verify that our fusion method is effective.

4. CONCLUSIONS

K-Means and KCCA for multimodal emotion recognition is proposed. The speech emotion feature extraction is time domain, frequency domain and MFCCs; the facial expression feature extracts the pixel point gray matrix. K-Means is used for feature clustering before fusion, and KCCA is used for serial fusion of feature matrices. SVM is used for feature recognition. Experimental results show that the recognition rate of this algorithm is higher than the rates procuded by other traditional fusion methods.

In future research, we will further explore the multimodal emotion recognition method at the feature level, find the inner relationship between modals, and apply it to the emotion recognition system to achieve efficient human-computer interaction.

REFERENCES

Chen, L.F., Feng, Y., Maram, M.A., Wang, Y.W., Wu, M., Hirota, K., and Pedrycz, W. (2019). Multi-svm based demp-ster-shafer theory for gesture intention understanding using sparse coding feature. *Applied Soft Computing*, DOI: 10.1016/j.asoc.2019.105787.

Chen, L.F., Wu, M., Zhou, M.T., Liu, Z.T., She, J.H., and Hirota, K. (2020). Dynamic emotion understanding in human-robot interaction based on two-layer fuzzy svr-ts model. *IEEE Transactions on Systems, Man, and Cybernetics Systems*, 50 (2), 490–501.

Chen, L.F., Zhou, M.T., Wu, M., She, J.H., Liu, Z.T., Dong, F.Y., and Hirota, K. (2018). Three-layer weighted fuzzy support vector regression for emotional intention understanding in human-robot interaction. *IEEE Transactions on Fuzzy Systems*, 26 (5), 2524–2538.

Ghahari, A., Fatmehsari, Y.R., and Zoroofi, R.A. (2009). A novel clustering-based feature extraction method for an automatic facial expression analysis system. In *5th International Conference on Intelligent Information Hiding and Multimedia Signal Processing, Kyoto, Japan*, 1314–1317.

Hotelling, H. (1935). Relations between two sets of variates. *Biometrika*, 28 (3/4), 321–377.

Kim, Y. and Provost, E.M. (2019). Isla: Temporal segmentation and labeling for audio-visual emotion recognition. *IEEE Transactions on Affective Computing*, 10 (2), 196–208.

Liu, C. and Wechsler, H. (2001). A shape- and texture-based enhanced fisher classifier for face recognition. *IEEE Transactions on Image Processing*, 10 (4), 598–608.

Melzer, T., Reiter, M., and Bischof, H. (2003). Appearance models based on kernel canonical correlation analysis. *Pattern Recognition*, 36 (9), 1961–1971.

Moore, B. (1981). Principal component analysis in linear systems: Controllability, observability, and model reduction. *IEEE Transactions on Automatic Control*, 26 (1), 17–32.

Nielsen, A. (2002). Multiset canonical correlations analysis and multispectral, truly multitemporal remote sensing data. *IEEE Transactions on Image Processing*, 11 (3), 293–305.

Peng, J., Li, Q., El-Latif, A., and Niu, X. (2015). Linear discriminant multi-set canonical correlations analysis (ldmcca): an efficient approach for feature fusion of finger biometrics. *Multimedia Tools and Applications*, 74 (13), 4469–4486.

Planet, S. and Sanz, I. (2012). Comparison between decision-level and feature-level fusion of acoustic and linguistic features for spontaneous emotion recognition. In *7th Iberian Conference on Information Systems and Technologies, Madrid, Spain*, 1–6.

Sultana, S. and Shahnaz, C. (2014). A non-hierarchical approach of speech emotion recognition based on enhanced wavelet coefficients and k-means clustering. In *3rd International Conference on Informatics, Electronics and Vision, Dhaka, Bangladesh*, 1–5.